Enhancing Transferable Adversarial Attacks on Vision Transformers through Gradient Normalization Scaling and High-Frequency Adaptation

Zhiyu Zhu, Xinyi Wang, Zhibo Jin, Jiayu Zhang, Huaming Chen

THE UNIVERSITY OF SYDNEY

ICLR

Introduction

- Traditional transfer attacks are very effective against CNN, but ViT attacks have limited effect.
- Mild gradients cause traditional attack methods to be ineffective against ViT.
- Combining Gradient Normalization Scaling (GNS) and High-Frequency Adaptation (HFA).

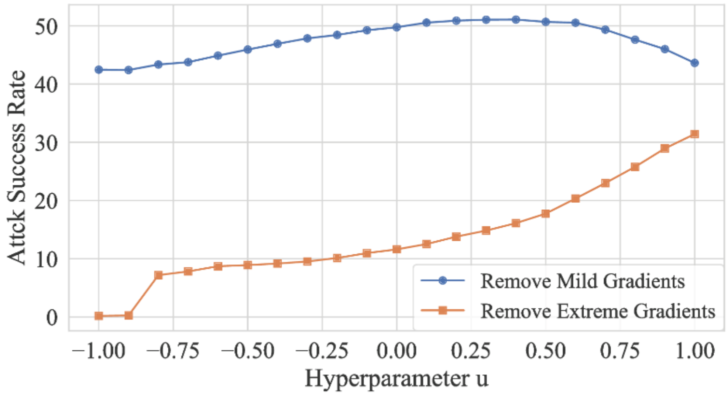
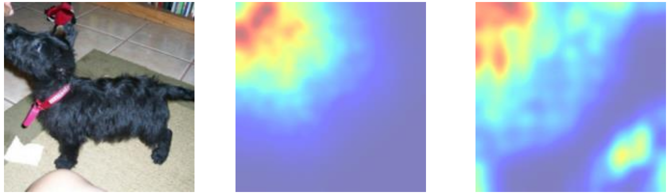


Figure 1: Performances of removing mild gradients and extreme gradients on ViT-B/16

Preliminaries

- Objective: Create perturbation η^$*$ that maximizes loss function $(L(f_\theta(x+\eta),y)$
- Constraint: Adhere to an L_p norm to ensure perturbation is imperceptible.
- Goal: Enhance perturbation effectiveness across multiple black-box models f_{θ_i}
- Mechanism: Multi-head Self-Attention (MSA) in ViTs.
- Performance: ViTs process different input sequence parts in separate spaces.
- Integration: Diverse informational perspectives integrated through Q, K, V matrices.
- Outcome: Robust model output via multiple attention calculations and projections.



Original Res-101 CaiT-S/24

Figure 2: Attribution visualization for different models in frequency domain.

GNS & HFA

- GNS (Gradient Normalizing Scaling):
    - Scales gradients based on deviation from mean:

$$g_l = g_l \cdot tanh\left(\left|\frac{g_l-\mu}{\sigma}\right|\right) \qquad (1)$$

    - g_l : Gradient for the l-th channel
    - μ, σ: Mean and standard deviation of gradients
    - Benefit: Normalizes gradient range, reduces overfitting
- HFA (High-Frequency Adjustment):
    - ViTs sensitive to high-frequency features
    - Gradient adaptation based on image frequency profile
    - Targets regions prioritized by ViTs
- High-Frequency Feature Exploration:
    - Uses mask to emphasize high-frequency areas: $mask_{ij}^{k} = \frac{\left(\frac{W+i}{2}\right)\cdot\left(\frac{H+j}{2}\right)}{W\times H}$ (2)
    - Frequency manipulation with DCT and IDCT

GNS-HFA Overview

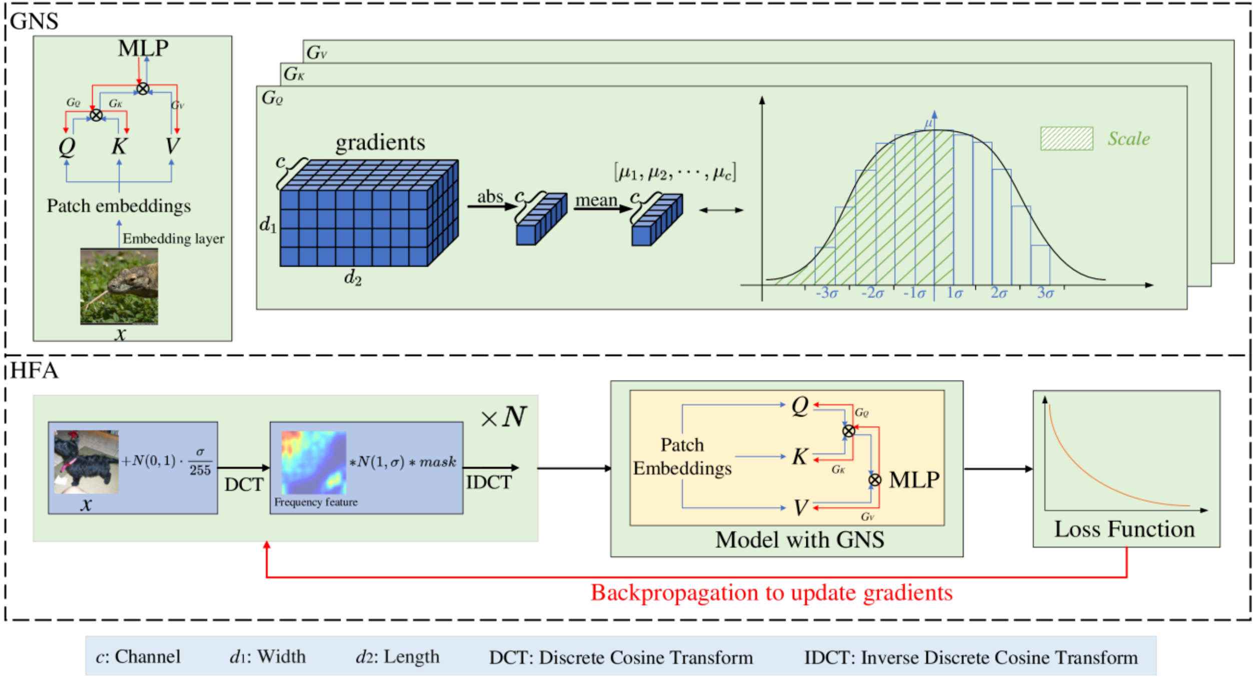


Figure 3: Illustration of our GNS-HFA method. Red links represent gradients backpropagation.

Experiments

- Dataset: 1000 images from the ILSVRC 2012 validation set
- ViT Models: LeViT-256, PiT-B~\citep, DeiT-B, ViT-B/16, TNT-S, ConViT-B, Visformer-S, and CaiT-S/24.
- CNN Models: Inception-v3, Inception-v4, Inception-ResNet-v2, and ResNet-101
- Metrics: Attack Success Rate (ASR)

Experiments

Surrogate Models	Method	ViT								CNN						
		LeViT-256	PiT-B	DeiT-B	ViT-B/16	TNT-S	ConViT-B	Visformer-S	CaiT-S/24	Inc-v3	Inc-v4	IncRes -v2	ResNet -101	Inc-v3 -adv-3	Inc-v3 -adv-4	IncRes -v2-adv
ViT-B/16	TGR	65.60%	55.70%	88.00%	99.60%	80.40%	88.40%	62.50%	86.60%	55.40%	50.60%	45.20%	51.30%	38.80%	38.40%	33.20%
	SSA	59.90%	59.20%	82.40%	99.80%	76.70%	83.70%	62.30%	83.40%	62.00%	59.60%	56.60%	58.20%	53.40%	53.90%	**50.80%**
	PNA	42.10%	41.80%	72.30%	94.00%	59.00%	71.30%	43.10%	71.70%	37.40%	35.30%	28.60%	33.80%	24.20%	23.40%	17.80%
	BIM	14.30%	14.80%	36.00%	100.00%	26.50%	39.10%	16.30%	38.20%	13.10%	10.60%	10.50%	11.60%	7.90%	5.90%	5.40%
	PGD	12.80%	12.50%	31.60%	100.00%	22.60%	34.10%	14.20%	33.20%	12.90%	10.10%	9.40%	12.80%	6.40%	4.30%	3.40%
	DI-FGSM	34.70%	37.50%	55.00%	98.30%	49.00%	59.60%	37.50%	58.70%	29.90%	30.00%	25.50%	26.70%	22.00%	21.40%	17.40%
	TI-FGSM	16.70%	19.90%	29.60%	97.40%	31.30%	34.40%	23.30%	30.50%	18.80%	18.60%	12.80%	16.90%	15.90%	17.40%	14.20%
	MI-FGSM	34.40%	33.90%	62.70%	99.90%	51.20%	64.20%	36.60%	64.70%	33.20%	30.50%	25.90%	32.30%	23.60%	21.10%	18.90%
	SINI-FGSM	45.70%	39.00%	75.30%	100.00%	65.80%	76.50%	45.10%	77.60%	46.00%	44.40%	36.50%	43.10%	36.60%	36.50%	31.10%
	GNS-HFA (Ours)	**76.80%**	**70.60%**	**93.50%**	99.80%	**87.60%**	**92.50%**	**72.70%**	**92.40%**	**67.30%**	**64.10%**	**59.00%**	**63.10%**	**54.50%**	**55.80%**	48.20%
Visformer-S	TGR	79.10%	71.50%	65.70%	43.50%	79.50%	58.00%	100.00%	67.80%	76.30%	75.90%	65.70%	72.40%	45.00%	38.90%	28.80%
	SSA	75.60%	73.70%	74.90%	64.10%	77.70%	73.80%	97.20%	75.40%	77.60%	76.90%	74.30%	74.90%	70.00%	69.30%	65.90%
	PNA	65.80%	61.90%	46.90%	28.80%	69.10%	44.40%	100.00%	52.40%	53.30%	53.20%	40.70%	45.70%	23.70%	19.90%	15.40%
	BIM	24.50%	27.20%	14.10%	9.40%	29.70%	16.70%	99.90%	16.10%	19.80%	19.40%	13.30%	16.40%	8.30%	6.60%	4.60%
	PGD	26.80%	24.20%	14.20%	10.90%	27.20%	14.60%	99.90%	15.10%	20.70%	20.90%	14.10%	17.20%	7.30%	5.80%	4.40%
	DI-FGSM	54.50%	56.00%	39.20%	22.30%	57.10%	39.60%	98.80%	45.10%	47.20%	47.90%	35.70%	39.40%	22.40%	17.30%	12.30%
	TI-FGSM	26.70%	34.70%	27.30%	19.90%	38.90%	29.60%	95.00%	29.30%	28.60%	28.00%	19.50%	21.80%	19.30%	21.80%	16.70%
	MI-FGSM	48.90%	50.70%	37.30%	29.30%	52.70%	38.90%	99.90%	40.50%	44.00%	43.20%	36.70%	39.30%	24.40%	21.50%	16.20%
	SINI-FGSM	68.00%	66.90%	58.30%	43.10%	72.00%	58.20%	100.00%	60.10%	63.50%	63.20%	55.00%	58.40%	40.30%	36.70%	30.10%
	GNS-HFA (Ours)	**94.90%**	**92.20%**	**91.10%**	**80.90%**	**94.60%**	**89.60%**	100.00%	**91.70%**	**95.30%**	**95.40%**	**92.70%**	**93.20%**	**89.60%**	**85.40%**	**80.20%**
PiT-B	TGR	87.80%	100.00%	83.20%	65.40%	90.50%	82.40%	88.50%	82.90%	80.00%	73.50%	69.30%	71.90%	51.10%	51.50%	40.50%
	SSA	64.20%	94.90%	66.90%	59.20%	71.00%	66.50%	67.10%	66.00%	63.80%	64.50%	59.30%	58.90%	55.20%	55.10%	51.80%
	PNA	62.20%	99.80%	54.60%	38.90%	67.00%	56.10%	70.50%	55.70%	51.40%	47.80%	41.80%	42.10%	25.70%	22.70%	16.60%
	BIM	17.60%	100.00%	11.80%	8.70%	23.50%	15.10%	22.20%	13.50%	16.30%	13.40%	10.70%	11.20%	6.90%	4.40%	3.60%
	PGD	17.00%	100.00%	11.10%	8.70%	20.10%	12.90%	20.20%	11.20%	14.90%	13.00%	11.90%	11.50%	5.90%	3.50%	3.40%
	DI-FGSM	43.60%	99.10%	38.80%	24.80%	54.10%	43.60%	56.40%	43.40%	36.70%	33.80%	26.50%	26.30%	16.20%	12.70%	9.60%
	TI-FGSM	21.50%	91.90%	24.80%	18.90%	32.70%	30.10%	35.20%	25.60%	20.60%	18.60%	13.60%	15.50%	14.60%	16.00%	11.90%
	MI-FGSM	38.10%	100.00%	34.30%	27.40%	46.70%	38.20%	44.60%	34.70%	35.90%	34.40%	27.10%	30.40%	19.10%	18.30%	14.00%
	SINI-FGSM	54.30%	100.00%	50.20%	37.60%	64.60%	52.10%	61.30%	53.30%	49.10%	46.80%	42.30%	44.10%	28.80%	28.10%	21.10%
	GNS-HFA (Ours)	**90.00%**	99.60%	**87.50%**	**75.50%**	**92.10%**	**87.80%**	**90.30%**	**86.60%**	**85.10%**	**82.00%**	**78.60%**	**78.80%**	**68.60%**	**70.20%**	**61.60%**
CaiT-S/24	TGR	82.70%	70.40%	98.80%	87.20%	93.50%	97.90%	81.30%	100.00%	68.60%	61.20%	59.40%	62.80%	49.10%	47.10%	38.30%
	SSA	77.30%	73.50%	88.40%	83.30%	87.70%	88.80%	77.30%	97.50%	75.60%	73.60%	72.60%	73.00%	69.10%	68.20%	**66.10%**
	PNA	59.70%	53.80%	82.70%	65.40%	76.20%	82.30%	59.50%	94.10%	49.20%	45.40%	41.70%	44.50%	31.80%	28.20%	22.90%
	BIM	26.70%	24.40%	73.90%	41.20%	51.90%	70.20%	30.30%	99.70%	20.30%	19.40%	15.40%	18.50%	10.50%	7.70%	6.00%
	PGD	25.70%	23.60%	67.80%	36.90%	45.00%	64.70%	27.20%	99.60%	20.70%	16.80%	15.70%	18.20%	7.30%	5.90%	4.70%
	DI-FGSM	60.80%	61.30%	83.30%	63.50%	78.40%	82.00%	64.80%	96.40%	51.70%	51.10%	46.80%	46.50%	34.10%	33.20%	27.30%
	TI-FGSM	36.20%	40.00%	61.10%	42.40%	59.00%	61.50%	47.90%	87.80%	30.40%	31.30%	24.00%	26.10%	26.80%	26.90%	22.60%
	MI-FGSM	54.80%	50.70%	90.20%	71.10%	78.80%	88.10%	55.50%	99.90%	48.70%	43.00%	39.50%	44.30%	31.60%	28.60%	23.30%
	SINI-FGSM	61.20%	53.80%	92.70%	77.50%	82.50%	92.10%	59.80%	100.00%	55.50%	50.40%	47.00%	50.70%	38.00%	38.30%	31.20%
	GNS-HFA (Ours)	**94.10%**	**87.70%**	**99.10%**	**95.90%**	**97.90%**	**98.90%**	**91.50%**	100.00%	**84.70%**	**80.70%**	**81.70%**	**81.90%**	**74.20%**	**73.60%**	64.60%

Table 1: ASR on ViT and CNN Models.

Conclusion

1. We enhance adversarial sample transferability by normalizing and scaling mild gradients during backpropagation, reducing overfitting.

2. We develop a HFA method to direct gradient updates more effectively, exploiting ViTs' sensitivity to high-frequency features.

3. GNS-HFA significantly boosts the transferability of adversarial attacks on ViTs.

4. Experiments shows a substantial improvement over existing methods, with gains of 33.54% for ViTs and 42.05% for CNNs. Our code is available at: https://github.com/LMBTough/GNS-HFA


Thanks you